The Efficacy of Unconditional Maximum Likelihood Bias Correction: Comment on Jansen, van den Wollenberg, and Wierda

Benjamin Wright
University of Chicago

Index terms: Item response theory, item parameter estimation; Item response theory, one-parameter logistic model; Maximum likelihood estimation, conditional; Maximum likelihood estimation, unconditional; Minimum chi-square estimation; One-parameter logistic model; Rasch model.

In the two previous articles, Jansen, van den Wollenberg, and Wierda (JVW) object to the bias in unconditional maximum likelihood estimation (UMLE) of Rasch parameters. Their comments on the necessity of the Rasch model for measurement are impeccable, as is their algebra. The practical consequences of their work, however, contradict their objections. The crucial question for practitioners is whether there is a convenient correction for UMLE bias which is accurate enough for practical purposes. Psychometricians can, in fact, find firm support for the use of UMLE in the articles by JVW.

Even though they perseverate on their discovery that the Wright-Douglas (1977) correction of $(K - 1)/K$ for UMLE bias (where K is the number of test items) is slightly inexact for very short tests, the conclusion JVW actually report is that the difference between bias predicted by $K/(K - 1)$ and the bias JVW observe "practically disappears" (van den Wollenberg, Wierda, & Jansen, 1988, p. 309) when tests have more than 10 items. This statement goes further than the Wright and Douglas (1977) recommendation of $(K - 1)/K$ for tests of more than 20 items.

UMLE bias has no effect on the relative position of items, and thus no effect on substantive interpretations of the variable defined by the item calibrations. There are, however, two applications in which bias in item difficulties could become a practical problem. These are the effects of item bias on person measurement and on test equating.

Person Measurement Bias?

What effect does UMLE item calibration bias have on person measures? The aim of testing is to provide person measures sufficiently accurate for fair evaluation. The $(K - 1)/K$ correction for bias is applied to the item difficulties after they are centered at 0. As a result, the measures most affected by error in bias correction are those associated with extreme scores $R = 1$ and $R = K - 1$ (where R is the

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 12, No. 3, September 1988, pp. 315–318
© Copyright 1988 Applied Psychological Measurement Inc.
0146-6216/88/030315-04$1.45

number of correct answers and K is the number of test items). To discover the maximum effect of the inaccuracy in $(K - 1)/K$ discussed by JVW on person measures, the values JVW claim to be "correct" (van den Wollenberg et al., 1988, Tables 1 through 3) and their associated item distributions and test lengths ($K = 5$, 10, 15, 20) were used to calculate the person measurement bias when $R = 1$.

The relevant UFORM (uniform approximation estimation) formulas, which are exact for the uniform tests used by JVW, are derived in Wright and Douglas (1975, pp. 21–24, 32) and applied in Wright and Stone (1979, pp. 143–151, 212–215). Table 1 provides the maximum person measurement bias both in logits (to show its tiny magnitude) and in standard error units (to show its statistical insignificance).

It is immediately clear that, even for $K = 5$, UMLE item bias is of no practical consequence as far as person measurement is concerned. Except for the 5-item, 8-logit test (a very rare configuration), maximum measurement bias is less than .21 logits (less than .16 standard errors of measurement).

For tests of usual length and width—more than 10 items, less than 6 logits—the maximum measurement bias due to JVW's results is *always* less than .09 logits (less than .08 standard errors of measurement). Even these minute discrepancies only occur when scores are extreme, $R = 1$ or $R = K - 1$. When tests are on target, observed scores cluster around $K/2$ where UMLE measurement bias is 0. It is clear that person measurement bias cannot be a reason to avoid UMLE.

Test Equating Bias?

What effect does UMLE item calibration bias have on test equating? In the Rasch method of equating two tests, a subset of common items is included in both, each test is calibrated separately, the resulting pairs of item estimates for the common items are plotted, and the intercept of a line with a slope of 1 fitted to these common item points is used as the equating constant (Wright & Stone, 1979, pp. 108–118).

In this procedure inaccuracy in $(K - 1)/K$ tends to cancel; this is especially true when tests are similar in length and difficulty (the usual situation), because then the inaccuracy is similar for the two calibrations. If, however, tests differ substantially in length and difficulty, then fitting a line with a slope adjusted to the distributions of common item difficulties can remove the effect of bias.

Table 1
Maximum Measurement Bias Due to UMLE Item Calibration After Correction $(K-1)/K$, in Logits and Standard Errors of Measurement

	Item Parameter Range			
K	−2,+2	−3,+1	−3,+3	−4,+4
Logits				
5	.05	.11	.20	.43
10	.02	.06	.09	.18
15	.02	.05	.06	.15
20	.02	.04	.05	.10
Standard Errors of Measurement				
5	.04	.09	.15	.31
10	.02	.05	.08	.15
15	.02	.05	.05	.14
20	.02	.04	.05	.09

The least biased and most efficient way to equate two or more tests linked by a network of common items and/or common persons is to combine the data from each administration into one large matrix, with a column for every item included in any test and a row for every person included in any sample, indicating missing data whenever a person does not answer an item. The single Rasch analysis of this one large matrix provides item calibrations and person measures on a common linear scale for all items and all persons involved in any test (Schulz, 1987; Wright & Linacre, 1985; Wright, Schulz, Congdon, & Rossner, 1987).

Conditional Estimation?

JVW advocate minimum chi-square pairwise estimation as their cure for the effects of $(K - 1)/K$ inaccuracy on UMLE. The logically equivalent but statistically superior maximum likelihood pairwise estimation method described by Rasch (1960/1980, pp. 171–172) and Choppin (1968) and applied extensively by Choppin (1976, 1977, 1978, 1983) is a better solution to this problem.

Rasch's pairwise method has significant antecedents in Case V of Thurstone's (1927a, 1927b) Law of Comparative Judgment, Bradley and Terry's (1952) method of paired comparisons, and Luce's (1959) probabilistic theory of choice. It is easy to use and understand, and generalizes directly to rating scale and partial credit models (Wright & Masters, 1982, pp. 67–72, 82–85). Should a real situation actually arise where conditional estimation is seriously deemed worth the trouble, then the Rasch/Choppin pairwise approach is the method of choice.

Conclusions

For practitioners working with tests of more than 10 items, the articles by Jansen, van den Wollenberg, and Wierda give no reason at all to avoid unconditional maximum likelihood estimation of Rasch item calibrations and person measures. In fact, their articles provide data which firmly support the adequacy of this practice.

References

Bradley, R. A., & Terry, M. E. (1952). Rank analysis of incomplete block designs. I. The method of paired comparisons. *Biometrika, 39*, 324–345.

Choppin, B. H. (1968). An item bank using sample-free calibration. *Nature, 219*, 870–872.

Choppin, B. H. (1976). Recent developments in item banking. In D. N. M. de Gruijter & L. J. T. van der Kamp (Eds.), *Advances in psychological and educational measurement*. New York: Wiley.

Choppin, B. H. (1977). Developments in item banking. In R. Sumner (Ed.), *Monitoring national standards of attainment in schools*. Windsor: National Foundation for Educational Research.

Choppin, B. H. (1978). *Item banking and the monitoring of achievement*. Slough: National Foundation for Educational Research.

Choppin, B. H. (1983). *A fully conditional estimation procedure for Rasch model parameters* (CSE Technical Report No. 196). Los Angeles: University of California. (ERIC Document No. ED 228267)

Jansen, P. G. W., van den Wollenberg, A. L., & Wierda, F. W. (1988). Correcting unconditional parameter estimates in the Rasch model for inconsistency. *Applied Psychological Measurement, 12*, 297–306.

Luce, R. D. (1959). *Individual choice behavior*. New York: Wiley.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research. [Expanded edition, University of Chicago Press, 1980.]

Schulz, E. M. (1987, April). *One-step vertical equating with MSCALE*. Paper presented at the Fourth International Workshop on Objective Measurement, University of Chicago, and the annual meeting of the American Educational Research Association, Washington.

Thurstone, L. L. (1927a). A law of comparative judgment. *Psychological Review, 34*, 273–286.

Thurstone, L. L. (1927b). The method of paired comparisons for social values. *Journal of Abnormal and Social Psychology, 21*, 384–400.

van den Wollenberg, A. L., Wierda, F. W., & Jansen, P. G. W. (1988). Consistency of Rasch model parameter estimation: A simulation study. *Applied Psychological Measurement*, *12*, 307–313.

Wright, B. D., & Douglas, G. A. (1975). *Best test design and self-tailored testing* (Research Memorandum No. 19). Chicago: Statistical Laboratory, Department of Education, University of Chicago.

Wright, B. D., & Douglas, G. A. (1977). Best procedures for sample-free item analysis. *Applied Psychological Measurement*, *1*, 281–294.

Wright, B. D., & Linacre, J. M. (1985). *MICROSCALE* [Computer program]. Westport CT: MEDIAX.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. Chicago: MESA Press.

Wright, B. D., Schulz, E. M., Congdon, R. T., & Rossner, M. (1987). *The MSCALE program for Rasch measurement*. Chicago: MESA Press.

Wright, B. D., & Stone, M. H. (1979). *Best test design*. Chicago: MESA Press.

Author's Address

Send requests for reprints or further information to Benjamin Wright, MESA Psychometric Laboratory, University of Chicago, Chicago IL 60637, U.S.A.